Classification Metrics

Anna Saranti

2026.05.03

1. Classification metrics
2. Mutual Information (MI)
3. Literature

1 Classification metrics

2 Mutual Information (MI)

3 Literature

Binary classification - Accuracy (1/3)

- Datasets are not perfect - therefore rarely balanced!

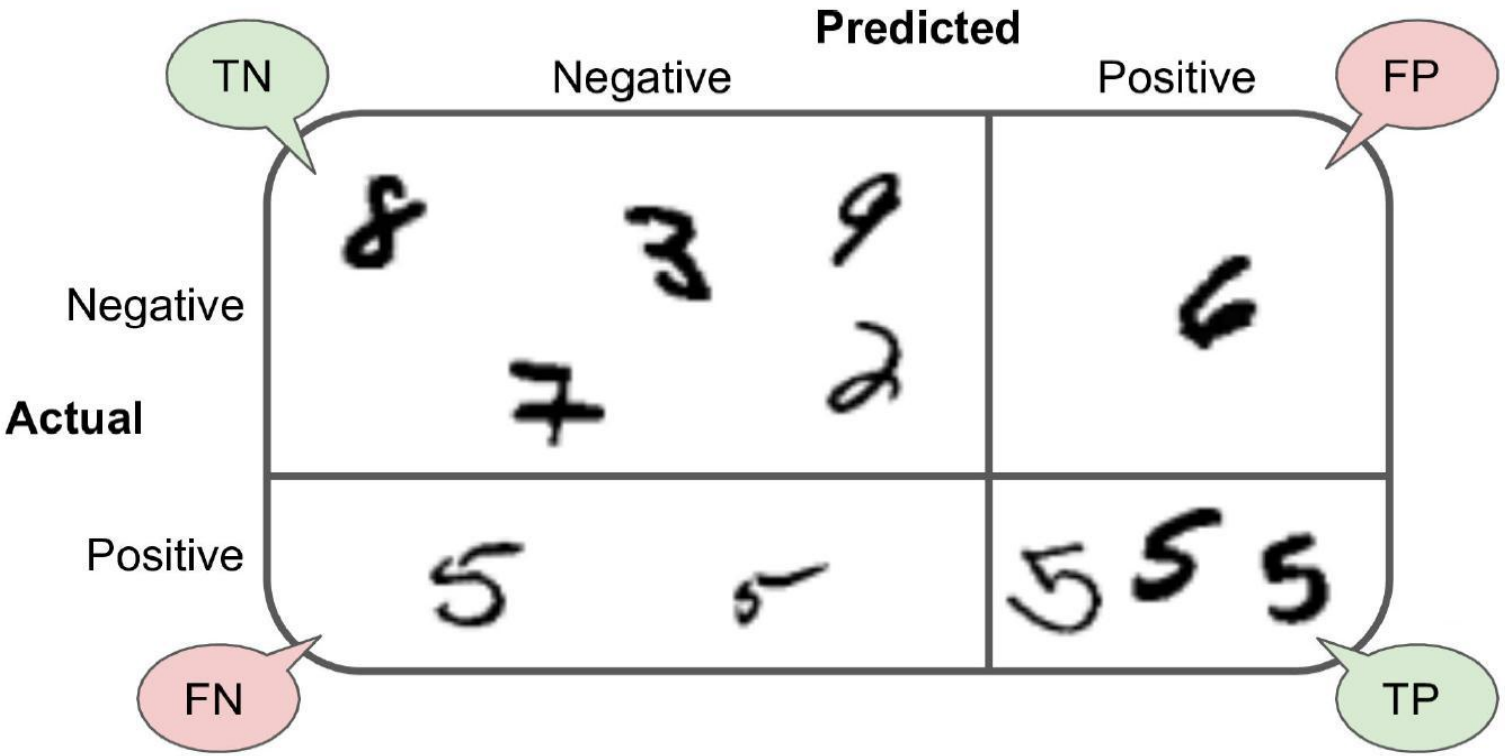
  Not just because of human error, but also because of phenomena that rarely happen

  Fault Detection - a fault happens (hopefully) rarely, a rare illness, ...

- accuracy $= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$

  TP: True positive, TN: True Negative
  FP: False positive, FN: False Negative

Binary classification - Accuracy (2/3)



Géron, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. "O'Reilly Media" 2019.

Binary classification - Accuracy (3/3)

```python
from sklearn.model_selection import StratifiedKFold
from sklearn.base import clone

skfolds = StratifiedKFold(n_splits=3, random_state=42)

for train_index, test_index in skfolds.split(X_train, y_train_5):
    clone_clf = clone(sgd_clf)
    X_train_folds = X_train[train_index]
    y_train_folds = y_train_5[train_index]
    X_test_fold = X_train[test_index]
    y_test_fold = y_train_5[test_index]

    clone_clf.fit(X_train_folds, y_train_folds)
    y_pred = clone_clf.predict(X_test_fold)
    n_correct = sum(y_pred == y_test_fold)
    print(n_correct / len(y_pred))  # prints 0.9502, 0.96565 and 0.96495
```

Géron, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. "O'Reilly Media" 2019.

Precision and Recall

- precision $= \frac{TP}{TP+FP}$, recall $= \frac{TP}{TP+FN}$
- precision-recall trade-off:

  For example: it is preferable to diagnose a person with cancer, even if the diagnosis is wrong then missing that a person indeed has cancer!

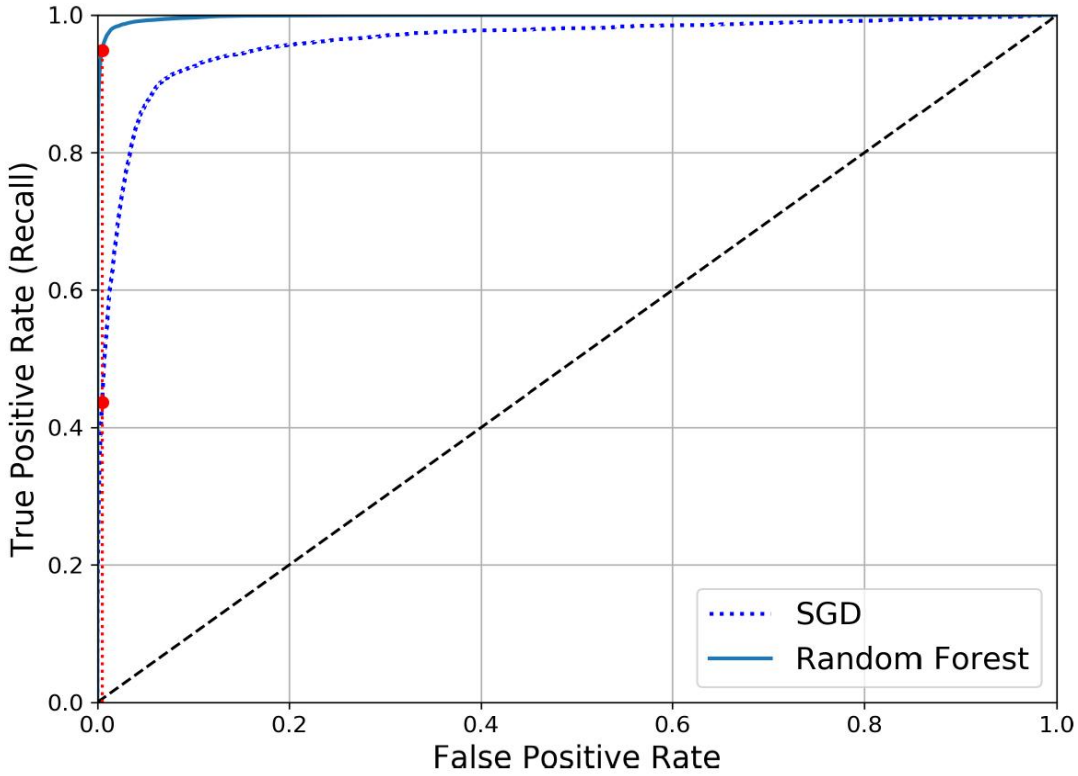
Receiver Operating Characteristic (ROC) (1/3)

- Receiver Operating Characteristic (ROC)
- Recall - True Positive Rate (TPR)
- Plot of True Positive Rate (TPR) against the
  False Positive Rate (FPR): ratio of negative instances that are incorrectly classified as positive -

  $1-$ True Negative Rate - specificity,
  which is the ratio of negative instances that are correctly classified as negative

- `sklearn.metrics.roc_curve`

Receiver Operating Characteristic (ROC) (2/3)



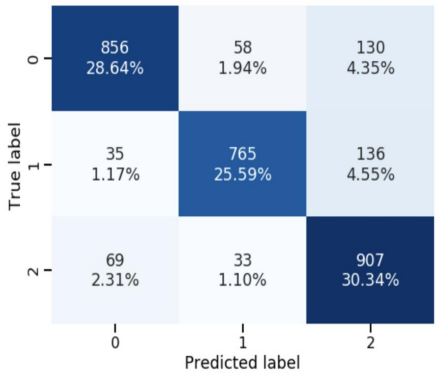
Géron, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. "O'Reilly Media" 2019.

Receiver Operating Characteristic (ROC) (3/3)

- Dotted line: ROC curve of a random classifier
- Good classifier - away from that dotted line as much as possible, near to the top-left corner
- Compare classifiers by comparing the area under the curve (AUC)

- `sklearn.metrics.roc_auc_score`

Multi-class classification (1/2)



https://medium.com/@dtuk81/confusion-matrix-visualization-fc31e3f30fea

- Accuracy can be computed, but we saw its problems
- Confusion matrix works here too
- BUT How can we compare f.e. a 2-class classifier with a 4-class classifier? Who is the best?

Multi-class classification (2/2)

- Scenario: virtual marketplace
  users sell and buy goods (items) privately
- The company does not know if the item has been bought successfully or not
- Let the user fill a form with 4 options:
  1. bought here
  2. bought elsewhere
  3. didn't buy here
  4. didn't buy anywhere else (no interest anymore)

1 Classification metrics

2 Mutual Information (MI)

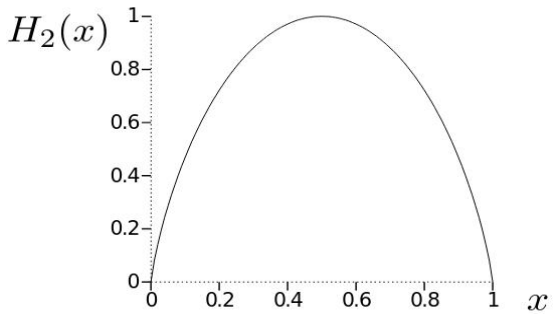
3 Literature

Entropy of a Random Variable (RV) (1/2)

- RV **X** with many outcomes,
  each of them represented by x
- Example: Throwing a dice -
  the set of possible outcomes: $1, 2, 3, 4, 5, 6$
  one possible outcome is: 5 $\leftarrow$ this is your x!
- Shannon information content of an outcome x
  $h(x) = \log_2 \frac{1}{P(x)}$ - measured in bits
- Entropy of ensemble **X** - Average Shannon information content of an outcome (also in bits):

  $H(\mathbf{X}) \equiv \sum_{x \in \mathcal{A}_\mathbf{X}} P(x) \log \frac{1}{P(x)}$
  $|\mathcal{A}_\mathbf{X}|$ number of elements in (finite) set $\mathcal{A}_\mathbf{X}$

Entropy of a Random Variable (RV) (2/2)



MacKay, David JC, and David JC Mac Kay. Information theory, inference and learning algorithms. Cambridge university press, 2003.

- $H((X)) \geq 1$ - equality iff $p_i = 1$ for one i
- $H((X)) \leq log(|\mathcal{A}_\mathbf{X}|)$ with equality iff $p_i = 1/|\mathcal{A}_\mathbf{X}|$ for all i
- Binary entropy $H_2(x) \equiv x\log\frac{1}{x} + (1-x)\log\frac{1}{(1-x)}$

Joint and Conditional Entropy (1/2)

- Joint ensembles in which Random Variables (RV) are dependent
- A noisy channel with input x and output y defines a joint ensemble in which x and y are dependent –
  if they were independent, it would be impossible to communicate over the channel

Joint and Conditional Entropy (2/2)

- Joint entropy H(**X**, **Y**) of **X**, **Y**:

  $\sum_{xy \in \mathcal{A}_{\mathbf{X}} \mathcal{A}_{\mathbf{Y}}} P(x, y) \log \frac{1}{P(x,y)}$

- Conditional entropy of **X** given **Y**:

  $H(\mathbf{X}|\mathbf{Y}) = \sum_{y \in \mathcal{A}_{\mathbf{Y}}} P(y) \left[\sum_{x \in \mathcal{A}_{\mathbf{X}}} P(x|y) \log \frac{1}{P(x|y)} \right] =$

  $\sum_{xy \in \mathcal{A}_{\mathbf{X}} \mathcal{A}_{\mathbf{Y}}} P(x, y) \log \frac{1}{P(x|y)}$

- Measures the average uncertainty about x, when y is known

Mutual Information (MI) (1/5)

- MI between $\mathbf{X}$ and $\mathbf{Y}$:
  $I(\mathbf{X}\,;\,\mathbf{Y}) \equiv H(\mathbf{X}) - H(\mathbf{X} \mid \mathbf{Y})$
- $I(\mathbf{X}\,;\,\mathbf{Y}) = I(\mathbf{Y}\,;\,\mathbf{X})$, $I(\mathbf{X}\,;\,\mathbf{Y}) \geq 0$
- Measures the average reduction in uncertainty about x that results from learning the value of y;

  or vice versa, the average amount of information that x conveys about y

Mutual Information (MI) (2/5)

- Input: $\mathbf{x}$, output of binary classifier: $y(\mathbf{x})$, target value t
- Error rate: fraction of misclassifications
- A and B have the same error rate of 10% and C has a greater error rate of 12% (in tables are percentages)

Classifier A

t \ y	0	1
0	90	0
1	10	0

Classifier B

t \ y	0	1
0	80	10
1	0	10

Classifier C

t \ y	0	1
0	78	12
1	0	10

MacKay, David JC, and David JC Mac Kay. Information theory, inference and learning algorithms. Cambridge university press, 2003.

Mutual Information (MI) (3/5)

- Classifier A simply guesses that the outcome is 0 for all cases and contains no information at all about t
- Classifier B has an informative output
  if $y = 0$ then we are sure that t really is 0
  if $y = 1$ then there is a 50% chance that $t = 1$
- Classifier C is slightly less informative than B, but it is still much more useful than the information-free classifier A

Mutual Information (MI) (4/5)

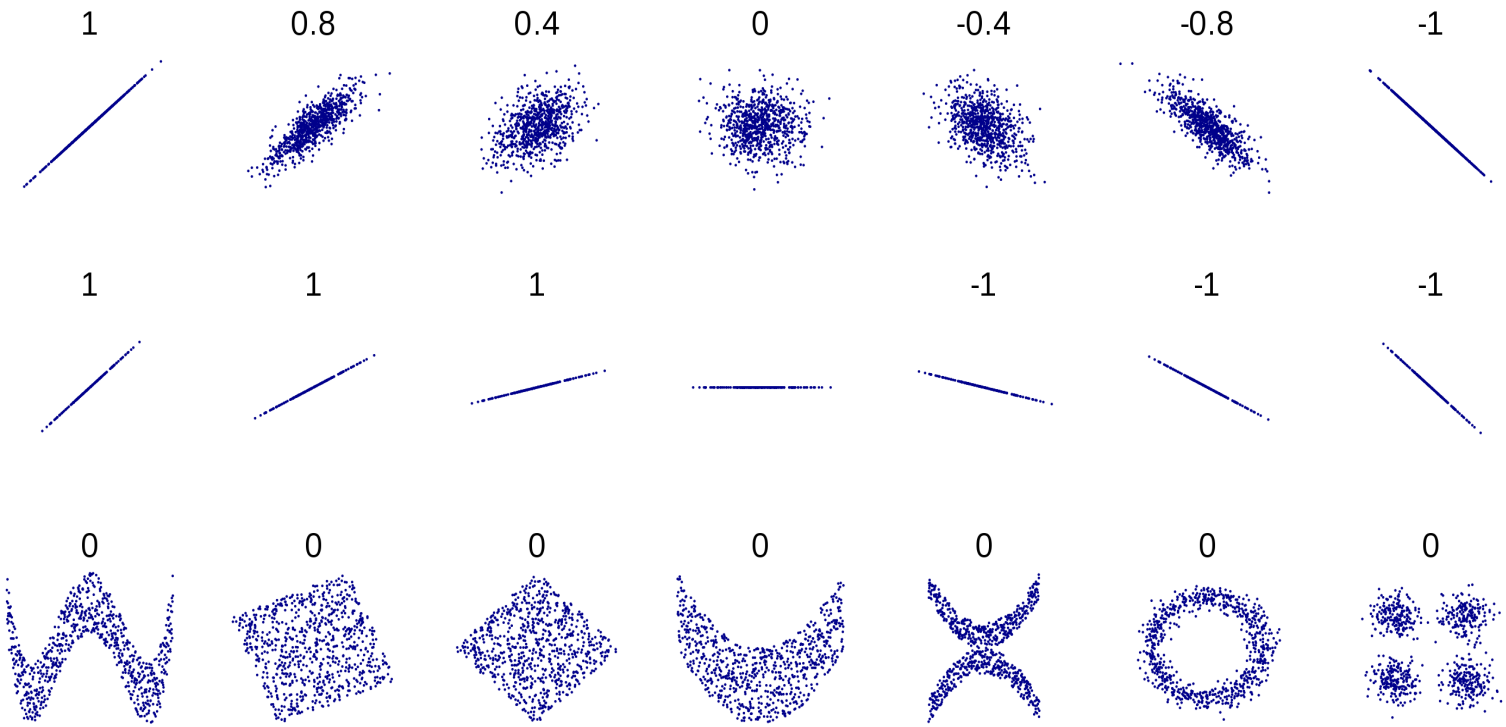
- Another thought: the error rate e is the sum of the two off-diagonal numbers
  the false positive rate e_+ and false negative rate e_-
- Classifier A's (e_+, e_-) is $(0, 0.1)$
  Classifier B's (e_+, e_-) is $(0.1, 0)$
  $\Longrightarrow$ you can't see immediately that classifier A is worthless!
- Need a performance measure that gives the worst possible score to classifier A -
  the error rate does not necessarily measure how informative a classifier' s output is

Mutual Information (MI) (5/5)

- Multi-class classification (digit recognition)
  the number of types of error increases from 2 to $10 \times 9 = 90$
  –one for each possible confusion of class t with t'
- We want one number!
- ROC: e_+ plotted versus e_-
- But wait a minute, e_+ and e_- are good?
  We've seen their deficiencies

  $\Longrightarrow$ "Does plotting one error rate as a function of another make this weakness of error rates go away?"

Correlations - Linear and non-linear (1/2)



https://en.wikipedia.org/wiki/Correlation

Correlations - Linear and non-linear (2/2)

- Top row: the correlation reflects the noisiness and direction of a linear relationship
- Middle row: Not the slope of the relationship
- Bottom row: Non-linear relationships

- Pearson correlation: `np.corrcoef`
- Mutual information: `sklearn.metrics.mutual_info_score`

1 Classification metrics

2 Mutual Information (MI)

3 Literature

Literature

- Géron, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. "O'Reilly Media" 2019.
- MacKay, David JC, and David JC Mac Kay. Information theory, inference and learning algorithms. Cambridge university press, 2003.